



King's Research Portal

DOI:

[10.1038/ng.3624](https://doi.org/10.1038/ng.3624)

Document Version

Peer reviewed version

[Link to publication record in King's Research Portal](#)

Citation for published version (APA):

Hore, V., Viñuela, A., Buil, A., Knight, J., McCarthy, M. I., Small, K., & Marchini, J. (2016). Tensor decomposition for multiple-tissue gene expression experiments. *Nature Genetics*, 48(9), 1094-1100.

<https://doi.org/10.1038/ng.3624>

Citing this paper

Please note that where the full-text provided on King's Research Portal is the Author Accepted Manuscript or Post-Print version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version for pagination, volume/issue, and date of publication details. And where the final published version is provided on the Research Portal, if citing you are again advised to check the publisher's website for any subsequent corrections.

General rights

Copyright and moral rights for the publications made accessible in the Research Portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognize and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the Research Portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the Research Portal

Take down policy

If you believe that this document breaches copyright please contact librarypure@kcl.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Tensor decomposition for multi-tissue gene expression experiments

Victoria Hore¹, Ana Viñuela², Alfonso Buil³, Julian Knight⁴, Mark I McCarthy^{4,5},
Kerrin Small², Jonathan Marchini^{1,4}

¹ Department of Statistics, University of Oxford, 24-29 St Giles, Oxford OX1 3LB, UK

² Department of Twin Research and Genetic Epidemiology, King's College London, SE1
7EH, UK.

³ Department of Genetic Medicine and Development, University of Geneva, Geneva,
Switzerland.

⁴ The Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK.

⁵ Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford,
Churchill Hospital, Old Road, Oxford OX3 7LJ.

Correspondence:

Professor Jonathan Marchini

Department of Statistics

University of Oxford

1 South Parks Road

Oxford OX1 3TG, UK

Tel: +44 (0)1865 271125

Fax: +44 (0)1865 281333

E-mail: marchini@stats.ox.ac.uk

35 **Abstract**

36

37 Genome wide association studies of gene expression traits and other cellular
38 phenotypes have been successful in revealing links between genetic variation
39 and biological processes. The majority of discoveries have uncovered *cis* eQTL
40 effects via mass univariate testing of SNPs against gene expression in *single*
41 tissues. We present a Bayesian method for multi-tissue experiments focusing on
42 uncovering gene networks linked to genetic variation. Our method decomposes
43 the 3D array (or tensor) of gene expression measurements into a set of latent
44 components. We identify sparse gene networks, which can then be tested for
45 association against genetic variation genome-wide. We apply our method to a
46 dataset of 845 individuals from the TwinsUK cohort with gene expression
47 measured via RNA sequencing in adipose, LCLs and skin. We uncover several
48 gene networks with a genetic basis and clear biological and statistical
49 significance. Extensions of this approach will allow integration of multi-omic,
50 environmental and phenotypic datasets.

51

52

53

54

55

56

57

58

59

60

61

62

63

64

65

66

67 **Introduction**

68

69 Studies of cellular phenotypes are transforming our understanding of the genetic
70 influences on complex traits. Genomic screens of gene expression levels¹,
71 chromatin accessibility², chromatin state³ and protein levels⁴ are all helping to
72 elucidate how genetics is related to disease mechanisms. Over the last few years
73 eQTL mapping has emerged as a key component in this research and has led to
74 the identification of many genetic variants affecting gene expression. Typically,
75 these studies involve assaying gene expression in a single tissue or cell type,
76 though multi-tissue experiments are beginning to emerge as a way to uncover
77 the principles of gene regulation.

78

79 The standard paradigm for *single tissue* eQTL studies involves testing the
80 expression of each gene or transcript against SNP genotypes in a local region to
81 identify *cis* eQTLs. This approach has been successful, with recent large eQTL
82 studies suggesting that there will be at least one *cis* eQTL for almost all
83 expressed genes⁵. Multi-tissue approaches can increase the power to find *cis*
84 eQTLs⁶, however, as *cis* eQTLs are estimated to account for only 30-40% of the
85 heritability of expression levels^{7,8} there is a need to identify *trans* eQTLs to
86 account for the remaining heritability.

87 The detection of *trans* eQTLs and networks of genes with related expression
88 patterns is hard both computationally and statistically. Testing all genes against
89 all SNPs via tens of thousands of genome-wide scans incurs a substantial penalty
90 for multiple testing. In addition, *trans* eQTL effect sizes tend to be smaller than
91 *cis* eQTLs making their detection harder⁹. For these reasons, scans for *trans*
92 eQTLs usually work with a reduced set of genetic variants, such as those
93 associated with disease traits^{9,10}. In general, the approach of carrying out very
94 large numbers of marginal statistical tests (of one SNP versus one gene at a time)
95 ignores the complex structure of these datasets. The expression levels of each
96 gene will likely be due to a mixture of several different sources, related to
97 underlying biology and also confounding factors.

98

99 In this paper we present a novel method for the analysis of multi-tissue gene
100 expression experiments, with a specific focus on identifying *trans* eQTLs and
101 gene networks. The data from such experiments can be viewed as a '3D' array, or
102 tensor, with dimensions representing individual, gene and tissue (see **Figure 1**).
103 Our method decomposes this tensor into a number of latent components (or
104 factors) that represent major modes of variation in the dataset. Each of these
105 components consists of three vectors of scores (or loadings) that indicate the
106 relative contribution of each individual, gene and tissue. For example, if a
107 consistent pattern of gene expression across a network of genes occurs in a
108 subset of tissues, with a different magnitude in each individual, then our model
109 aims to represent this in a single component. Such signals might naturally arise
110 due to transcription factor genes that have multiple targets throughout the
111 genome. If the expression level or function of a gene is altered by *cis*-acting
112 genetic variants, then we would likely observe different magnitudes of effects
113 across individuals.

114
115 One useful way to think about the approach is as analogous to the use of
116 principal components analysis (PCA) applied to '2D' (individual by SNP) genetic
117 datasets. PCA is routinely used to decompose genome-wide SNP datasets into
118 components of variation that are then used to understand population structure
119 (see ¹¹ for example). Here we aim to decompose higher dimensional datasets into
120 components that uncover real biology.

121
122 Our method has several notable properties:

- 123 • Our approach is developed in a Bayesian framework, and we use a sparse
124 'spike and slab'¹² prior to allow the gene loadings of each component to
125 have a unique level of sparsity. This allows us to shrink gene effects to
126 zero so that we can infer more clearly which genes are involved in gene
127 networks.
- 128 • The individual scores represent the magnitude of the effect of each
129 component across individuals, analogously to the individual scores that
130 are usually plotted in a PCA analysis of genetic datasets. We use these
131 scores as phenotypes in genome-wide SNP scans to identify genetic

132 variants that drive each component. The number of components we test is
133 typically much smaller (a few hundred) than the number of genes (tens of
134 thousands), which substantially reduces the multiple testing burden
135 when compared to approaches that test all genes against all genetic
136 variants in all tissues.

- 137 • We do not claim that all genes identified in a network will reach genome-
138 wide significance thresholds with the driving SNPs. However, when
139 applied to real datasets we find that the majority of genes are nominally
140 significant.
- 141 • The tissue scores vector indicates the ‘activity’ of the component for each
142 tissue. By examining the entries of the tissue scores matrix across
143 components we can make inferences about how many components are
144 shared across tissues.
- 145 • Our model also allows for non-sparse components that might be expected
146 to arise from confounding effects, such as batch effects or sequencing
147 properties.
- 148 • In addition, the model can naturally accommodate missing data, such as
149 samples without gene expression on subsets of tissues, which is a real and
150 prevalent feature of multi-tissue experiments.

151
152 Our motivation for this work stemmed from similar approaches that have
153 emerged in the field of neuroscience to uncover shared signals across different
154 high-dimensional imaging modalities^{13,14}. Most tensor decomposition methods¹⁵⁻
155 ¹⁷ are not able to handle missing data or invoke sparsity on the components,
156 although there are some exceptions¹⁸. Our model is the first tensor
157 decomposition method utilizing spike and slab priors with model fitting carried
158 out using Variational Bayes (see **Online Methods**). Via extensive simulations
159 (**Supplementary Note**) we show that our method has the best performance in
160 terms of estimation of the component individual scores and recovery of sparsity
161 patterns in the gene loadings when compared to other matrix and tensor
162 decomposition methods, and is well powered to detect *trans* eQTL signals and
163 gene networks. Our method is implemented in a software package called **SDA**
164 (Sparse Decomposition of Arrays) (see URLs).

165

166 **Results**

167

168 We have validated our approach by applying it to RNA sequencing data from the
169 TwinsUK cohort, which consists of gene expression measured on 845 related
170 individuals in adipose, LCLs and skin^{19,20}. In order to focus on robustly identified
171 components we applied our method 10 times to the TwinsUK RNA-seq dataset
172 and combined results across runs via clustering (see **Online Methods**). After
173 clustering, we identified 236 robust components for further investigation.
174 Examination of the tissue scores matrix is informative about which tissues each
175 component is active in (see **Supplementary Figure 1**). We found that the
176 majority of the 236 components were active in a single tissue (57 in Adipose, 74
177 in LCLs and 70 in Skin). There were 20 components that were active in all 3
178 tissues, 14 components active in just Adipose and Skin and 1 component active
179 in Adipose and LCLs. The full details of these 236 components are given in the
180 **Supplementary Data Set**.

181

182 The individual scores vectors of these components were then used as
183 phenotypes in genome-wide scans using SNP genotype data imputed using the
184 1000 Genomes Phase 1 reference panel. We used a threshold of 1×10^{-10} to assign
185 significance (see **Online Methods**). There were 26 components that reached this
186 level of significance: 5 were active in just 1 tissue (1 in Adipose, 4 in LCLs), 20
187 components were active in all 3 tissues and 1 component was active in just
188 Adipose and Skin. The majority of these components were clearly uncovering *cis*
189 eQTLs. In all but two of these components we identified pairs of SNPs
190 (significantly associated with our component scores) and genes (with a non-zero
191 loading) that had previously been identified as a *cis* eQTLs in the MuTHER and
192 GTEx studies^{7,21}. These components exhibited very sparse gene loadings, with a
193 single localized cluster of high gene loadings and highly significant SNP
194 associations in the flanking region (**Supplementary Figures 2-21**).
195 Methodology for the detection of *cis* eQTLs is well established and is best carried
196 out using focused analysis that looks for such effects at SNPs flanking each gene.

197 Our main focus is on uncovering *trans* eQTLs and gene networks so we do not
198 pursue the *cis* eQTLs that our method finds any further.

199

200 The remaining 6 components were less sparse in their gene loadings, and
201 exhibited patterns of gene loadings and SNP associations that highlight gene
202 expression networks with substantial biological significance. For these networks
203 the majority of gene loadings tend to be unidirectional suggesting the
204 components are identifying a directional effect on expression. These components
205 are summarized in **Figures 2-6** which show the gene loadings, SNP GWAS and
206 tissue activation patterns. **Supplementary Table 1** shows that the majority of
207 genes identified in each of these networks have nominally significant p-values in
208 the relevant tissues. At the suggestion of a reviewer, we also applied PCA and ICA
209 to the Twins UK dataset. Neither of these approaches uncovered the gene
210 networks reported here; more details are given in the **Supplementary Note**.

211

212 We found 2 clustered components (**Figure 2**) with individual scores that exhibit
213 significant SNP associations in the gene *CIITA* on chromosome 16p13 (see also
214 **Supplementary Figure 22**). The first component is active mostly in Adipose and
215 Skin and has a lead SNP rs9924520 (p-value = 1.33×10^{-23} , MAF=0.247) that is an
216 intronic variant of *CIITA*. The second component is active mostly in LCLs and has
217 a lead SNP rs7194862 (p-value = 1.74×10^{-14} , MAF=0.282) that is 5' of *CIITA*. The
218 SNPs rs9924520 and rs7194862 are in strong LD ($r^2 = 0.82$). Both components
219 show a cluster of MHC Class II genes on chromosome 6 with non-zero gene
220 loadings. In addition, 2 other genes have significant gene loadings in both
221 components (*RFX5* on chromosome 1 and *CD74* on chromosome 5). *CIITA* is
222 known to be a master controller in the regulation of MHC Class II gene
223 expression²². It is recruited to the proximal promoter regions of the classical
224 MHC class II genes (*HLA-DP*, *HLA-DR* and *HLA-DQ*), and to *HLA-DM*, *HLA-DO* and
225 the *CD74* gene (encoding the molecular chaperone invariant chain which
226 associates with the MHC class II complex) through protein-protein interactions
227 with other components of the MHC class II enhanceosome, which includes *RFX5*.
228 **Supplementary Table 2** details the direct associations of SNPs rs9924520 and
229 rs7194862 with the expression levels of all the genes identified in our

230 components (in all three tissues) after correction for covariates and 15 PEER
231 factors²³ (see **Online Methods**). Both SNPs are strongly associated with *HLA-*
232 *DOA* and *HLA-DMA* in Adipose and Skin (p-values in the range [2.89×10^{-8} ,
233 5.56×10^{-19}]) and with *CIITA* in Adipose (p-values = 2.08×10^{-11} , 1.44×10^{-12}).
234 However, neither SNP reaches a strict Bonferroni threshold for a *trans* analysis
235 of $9.05 \times 10^{-13} = 5 \times 10^{-8} / (3 \times 18409)$ (obtained by accounting for genome-wide
236 testing across all genes in all tissues) with any of the other genes in the 3 tissues.
237 These results suggest that while a *trans* eQTL association would have been found
238 between SNPs in the *CIITA* region and expression at two MHC class II genes, the
239 more extensive network of genes recovered by our components would not have
240 been uncovered via a marginal *trans* analysis.

241

242 **Figure 3** shows significant associations in the gene *NLRC5/CITA* on chromosome
243 16q13 (see also **Supplementary Figure 23**). The lead SNP rs289749 (p-value =
244 1.34×10^{-11} , MAF=0.3) is an intronic variant of *NLRC5/CITA*. The component
245 shows a cluster of genes on chromosome 6 with non-zero gene loadings that
246 include MHC Class I genes (*HLA-O*, *HLA-B*, *HLA-F*, *HLA-A*, *HLA-E*), *BTN* genes
247 (*BTN3A2*, *BTN3A1*, *BTN3A3*, *BTN2A2*, *BTN2A1*), *TAP1*, *TAP2*, *PSMB8* and
248 *PSMB9*. Overexpression of *NLRC5/CITA* is known to increase mRNA levels of
249 genes encoding human MHC Class I molecules and proteins functioning in the
250 MHC Class I mediated antigen presentation pathway, including beta-2-
251 microglobulin (*B2M*), transporter associated with antigen processing 1 (*TAP1*)
252 and the proteasome subunit beta type-9 (*PSMB9*)²⁴. *B2M*, *TAP1* and *PSMB9* all
253 have significant gene loadings in the component. **Supplementary Table 3**
254 details the direct associations of SNP rs289749 with the expression levels of all
255 the genes in the component in all three tissues. In skin, rs289749 is strongly
256 associated with *NLRC5/CITA* (p-value = 1.37×10^{-28}) and moderately associated
257 with several MHC class I genes; *HLA-F* (p-value = 3.02×10^{-12}), *HLA-A* (p-value =
258 1.22×10^{-9}) and *HLA-B* (p-value 1.35×10^{-10}); although none of these associations
259 would pass a Bonferroni corrected significance level in a *trans* analysis (9.05×10^{-13}).
260 p-values for association between rs289749 and other genes in this
261 component suggest that the link between *NLRC5/CITA* and *BTN*, *TAP* and *PSMB*
262 genes or the *B2M* gene would not have been recovered using a traditional *trans*

analysis. In addition, these direct associations fail to provide evidence for the signal in either Adipose or LCLs.

Figure 4 shows significant associations for a cluster of SNPs near *LSM11* on chromosome 5q33.3 which is known to be involved in histone RNA processing²⁵ (see also **Supplementary Figure 24**). The gene loadings of our component show a striking cluster of 23 histone genes in the chromosome 6p21 cluster as well as the gene *HIST2H2BE* in the 1q21 cluster (**Figure 4** purple points). There are also additional signals at other histone genes on chromosome 1q42 (*HIST3H2A*), 11q23 (*H2AFX*) and 12p12 (*HIST4H4*). SNP rs6882516 (p-value = 2.39×10^{-15} , MAF=0.206) is in the 3' UTR of *LSM11* and predicted to be a microRNA binding site using mirSNP²⁶. Key histone gene regulatory factors are organized in a limited number of subnuclear foci. It is known that cell cycle-dependent phosphorylation of p220^{NPAT} by cyclin E/CDK2, that induces histone gene transcription, occur at these intranuclear sites. p220^{NPAT} colocalizes with both (a) the histone gene clusters on chromosome 1q21 and 6p21, (b) the protein subunit *LSM11*¹³. A set of 31 significant genes (loadings with a PIP>0.5, see **Online Methods**) show Gene Ontology p-values of 1.91×10^{-25} and 1.40×10^{-24} for the terms 'nucleosome organization' and 'chromatin assembly or disassembly' respectively. The tissue scores indicate that this component is only active in LCLs. **Supplementary Table 4** details the direct associations of SNP rs6882516 with expression levels of *LSM11* and the other genes in this component in all three tissues. The SNP is significantly associated with *LSM11* in LCLs (p-value = 5.57×10^{-33}), and has p-values in the range (2.65×10^{-12} , 1.17×10^{-12}) with three histone genes in our component with extreme gene loadings (*HIST1H1C*, *HIST1H2BJ* and *HIST1H2BK*). Although these associations are encouraging, they do not pass a strict *trans* analysis significance level and additionally, these direct associations do not uncover the link between *LSM11* and the histone gene cluster on 1q21 (the p-value for rs6882516 and *HIST2H2BE* in LCLs is 5.40×10^{-9}).

Figure 5 shows significant associations near the gene *USP18* (see also **Supplementary Figure 25**). The lead SNP rs2401506 (p-value = 9.82×10^{-16} ,

295 MAF=0.358) is 5kb upstream of *USP18*. The set of 160 genes in the loadings with
 296 a PIP>0.5 show Gene Ontology p-values of 1.73×10^{-42} and 1.23×10^{-38} for the
 297 terms 'defense response to virus' and 'response to type I interferon' respectively.
 298 Of the 70 genes annotated by 'response to type I interferon' we find 28 with non-
 299 zero gene loadings (**Supplementary Figure 26**). These include all four of the 2'-
 300 5' oligoadenylate synthetase (OAS) gene family (*OAS1*, *OAS2*, *OAS3* and *OASL*)
 301 known to be actively induced by interferons²⁷, the genes *STAT1* and *STAT2*
 302 which are key mediators of type I and type III IFN signaling, several Interferon γ -
 303 inducible protein (IFI) genes (*IFI6*, *IFI44L*, *IFI16*, *IFIH1*, *IFIT1*, *IFIT3*, *IFIT5*, *IFIT2*,
 304 *IFITM1*, *IFITM2*, *IFI35*) and the genes *MX1* and *MX2* also related to IFN signaling.
 305 *USP18*, a type I IFN-induced protein that deconjugates the ubiquitin-like modifier
 306 *ISG15* (which is also in our component) from target proteins²⁸, plays an
 307 important function in down regulation of interferon responses^{29,30} and
 308 significantly inhibits tumour growth³¹. The tissue scores indicate that this
 309 component is only active in LCLs. **Supplementary Table 5** details the direct
 310 associations of SNP rs2401506 with the 160 genes identified in this component
 311 across all three tissues. There is only evidence of association in LCLs, with
 312 several genes obtaining p-values smaller than 1×10^{-8} (*IFIT1*, *PLSCR1*, *STAT1*,
 313 *CMPK2*, *RSAD2* and *EIF2AK2*) but none are significant when accounting for
 314 genome-wide testing across all genes, suggesting that this network of genes
 315 would not have been uncovered by a scan of all SNPs versus all genes.

316
 317 **Figure 6** shows two significant associations on separate chromosomes for a
 318 component with a striking cluster of non-zero gene loadings for zinc finger genes
 319 on chromosome 19. SNP rs17611866 (p-value = 5.40×10^{-21} , MAF = 0.251) on
 320 chromosome 16 is a mis-sense variant in *ZNF75A*, which is one of 6 ZNF genes in
 321 a local cluster. Flanking genes *ZNF263* and *TIGD7* have non-zero gene loadings
 322 (see **Supplementary Figure 27**). SNP rs12630796 (p-value = 5.10×10^{-17} , MAF =
 323 0.487) on chromosome 3 is an intronic SNP in *SENP7*. A SNP in high LD with this
 324 SNP (rs13320918, p-value = 7.34×10^{-15} , MAF = 0.377) has been shown to be a
 325 microRNA QTL for miR-1270 (p-value= 1.71×10^{-10}) which is located in a zinc
 326 finger cluster on chromosome 19p12³². In a separate study, 4 other intronic
 327 SNPs in *SENP7* (rs2553419, rs2682386, rs9859077 and rs2141180), all in high

LD with each other and with rs13320918, were shown to correlate with *cis*-acting regulation of *SENP7* expression in CD4 and CD8 lymphocytes and *trans*-acting regulation of *ZNF154*, *ZNF274* and *ZNF814*³³, which all reside within a ~250-kb region on chromosome 19q13.43 (see **Supplementary Figure 28**).

Supplementary Table 6 details the direct associations of SNPs rs12630796 and rs17611866 with *SENP7* on chromosome 3 and genes with non-zero gene loadings in the component in all three tissues. This analysis partially recovers the signal that we find using our method, see the **Supplementary Note** for more details.

It can be challenging to interpret the large number of components that are produced by sparse matrix and tensor decomposition methods. By clustering components across independent runs of the method, and then selecting components with genetic associations, we have shown that it is possible to identify gene networks with clear biological significance. However, we have found evidence that the components without genetic associations are also capturing important variance in the data. Many components have individual scores vectors that are significantly associated with variables measuring properties of the sequencing; these components are mostly dense with several thousand non-zero gene loadings (see **Supplementary Figures 29-31** and **Supplementary Table 7**). Similarly, we have identified several components that are significantly associated with measured phenotypes including age, BMI and cholesterol levels (**Supplementary Figure 32**). We find two components that show association with age. These components are shown in **Supplementary Figures 33 and 34**. The most significant molecular function ontology term for both components is 'oxidoreductase activity' with p-values of 1.9×10^{-24} and 2.1×10^{-22} .

In addition, we have found that it can be useful to examine the components from a single run of the method. Specifically, we focus on the best run of 10 that produces the highest value of the model negative free energy (**Online Methods**).

361 We identified all components highlighted in **Figures 2-6** with significant or very
362 close to significant GWAS p-values. In addition, we find several components that
363 identify *KLF14* as a master *trans* regulator³⁴ (for example, see **Supplementary**
364 **Figure 35**). More details are given in the **Supplementary Note** and the
365 **Supplementary Data Set**.

366

367 A previous analysis of a similar set of samples in the MuTHER study⁷ using a
368 microarray based gene expression experiment called 518, 491 and 493 *trans*
369 eQTLs SNPs at a normal GWAS threshold of 5×10^{-8} . They reported an FDR of <
370 10% at this threshold, however only ~5% of these signals replicated at a
371 nominal significance threshold of 0.05 in at least one out of 5 other studies. The
372 overlap with our results is (a) a SNP rs7714390 on chromosome 5 (near our lead
373 SNP rs6882516) associated with two Histone genes (HIST1H2BK on chr 6 with a
374 p-value = 8×10^{-9} in LCLs and HIST2H2BE on chr1 with a p-value of 3.2×10^{-8} in
375 LCLs) (b) a SNP rs220377 on chr 16 (near our lead SNP rs17611866) associated
376 with a Zinc finger gene (ZNF667 on chr 19 with a p-value = 2.9×10^{-9} in LCLs), and
377 (c) several associated SNPs near rs4731702 that overlap with the KLF14
378 network with p-values between 4.4×10^{-8} and 2.2×10^{-15}). This analysis did not
379 identify the Type I Interferon network or the MHC networks that we find in our
380 analysis.

381

382

383 **Discussion**

384

385 We have described a new algorithm for efficient tensor decomposition for multi-
386 tissue gene expression datasets, and have demonstrated its utility on a real, three
387 tissue dataset to uncover sparse gene networks with clear biological and
388 statistical significance. A marginal analysis of all SNPs versus all genes would not
389 have uncovered these networks in the same way or with as much power. For
390 example, no aspect of the Type I interferon component would have been
391 identified. We have further shown in simulations that our method has good
392 power to detect sparse gene networks correlated to genetic variants, and dense
393 confounding factors.

394 This approach complements current eQTL analysis pipelines that tend to mainly
395 focus on identifying *cis* eQTLs in one tissue at a time. Analysis of cross tissue
396 effects usually proceeds in a subsequent step by comparing effect sizes across
397 tissues. Our method focuses on decomposing the complete multi-tissue dataset
398 into components of variance with varying levels of sparsity. We then test each
399 component against genetic variation genome-wide to uncover underlying eQTL
400 effects, ensuring robustness by only considering components that are
401 consistently found across multiple runs. We view our approach as
402 complementary to an association analysis of all SNPs versus all genes, since it
403 requires 2 orders of magnitude fewer tests, and has more power to detect SNP
404 associations with gene networks.

405

406 In general, we find that dense components uncovered by our method show high
407 levels of significance with confounding variables and the method additionally
408 uncovers many very sparse components that represent *cis* eQTLs. More
409 interestingly, we find 6 components with intermediate levels of sparsity with
410 gene loadings spread across multiple chromosomes that represent gene
411 networks showing a highly significant association with genetic variants. In all 6
412 of these components, we are able to link the gene networks they describe to
413 known biology. In the future it will be natural to apply this method to gene
414 expression datasets with even more tissues, such as that being collected by the
415 GTEx Project³⁷ or the Allen Institute for Brain Science (AIBS) human microarray
416 data set³⁸.

417

418 There are several interesting ways in which this model can be extended or
419 changed. The method can be naturally extended to higher dimensional datasets.
420 For example, 4D multi-tissue gene expression experiments through time and/or
421 under different experimental conditions (see **Supplementary Figure 36**).

422

423 One assumption of our model is that the gene loadings pattern of a component is
424 constant across active tissues, which may or may not be true dependent upon the
425 dataset being analyzed. One way to overcome this would be to develop a model
426 that applies a matrix decomposition to the gene expression matrix for each

tissue but with a linked individual scores matrix (see **Supplementary Figure 37**). A downside of such an approach is that it would significantly increase the number of unknown parameters in the factorization. However, this model would allow variation in the gene loadings between tissues if there were indeed clear differences, and might be a way of combining together components found by our tensor method (like those describing MHC class II regulation pathways) with clearly similar gene loadings. However, it may also be necessary to model the similarity between gene loadings to aid estimation, given the larger parameter space. This approach has strong connections to sparse canonical correlations analysis (CCA)³⁹ and unsupervised multi-view learning⁴⁰.

437

438

Such a linked matrix decomposition method could also be used to integrate different genomic datasets. The model has no constraint that the set of matrices being jointly decomposed have the same dimensions. So, for example, matrices of gene expression and epigenetic measurements could be jointly decomposed to uncover relevant shared biology (see **Figure 7**). Example applications might include joint decomposition of different omics datasets collected on cancer samples from the International Cancer Genome Consortium (ICGC) (see **URLs**). This model can further be extended to tensors of different data types (see **Supplementary Figure 38**).

448

449 **URLs**

450 SDA Software : <http://www.stats.ox.ac.uk/~marchini/sda.html>

451 ICGC <https://dcc.icgc.org/projects/details>

452 **Acknowledgements**

453

We are grateful to Andrew Dahl, Warren Kretzschmar, Kevin Sharp, Lloyd Elliot and Simon Myers for helpful discussions about the method and interpretation of the results. The TwinsUK cohort was funded by the Wellcome Trust and the European Community's Seventh Framework Programme (FP7/2007-2013). The

study also receives support from the National Institute for Health Research (NIHR) Clinical Research Facility at Guy's & St Thomas' NHS Foundation Trust and NIHR Biomedical Research Centre based at Guy's and St Thomas' NHS Foundation Trust and King's College London. SNP Genotyping was performed by The Wellcome Trust Sanger Institute and National Eye Institute via NIH/CIDR. Ana Viñuela and Alfonso Buil were supported by the EU FP7 grant EuroBATS (No. 259749). Victoria Hore acknowledges EPSRC for funding through a studentship at the Life Sciences Interface program of the University of Oxford's Doctoral Training Center. Jonathan Marchini acknowledges support from the ERC (Grant no. 617306).

Author Contributions

V.H and J.M developed the method. V.H carried out all analysis. J.M and V.H wrote the paper. A.V, A.B and K.S provided the TwinsUK dataset. A.V, A.B, J.K, M.M and K.S advised on interpretation of the results.

References

1. Stranger, B. E. *et al.* Population genomics of human gene expression. *Nat. Genet.* **39**, 1217–1224 (2007).
2. Degner, J. F. *et al.* DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature* **482**, 390–394 (2012).
3. Kasowski, M. *et al.* Extensive variation in chromatin states across humans. *Science* **342**, 750–752 (2013).
4. Battle, A. *et al.* Genomic variation. Impact of regulatory variation from RNA to protein. *Science* **347**, 664–667 (2015).
5. Pai, A. A., Pritchard, J. K. & Gilad, Y. The genetic and mechanistic basis for variation in gene regulation. *PLoS Genet.* **11**, e1004857 (2015).
6. Flutre, T., Wen, X., Pritchard, J. & Stephens, M. A statistical framework for joint eQTL analysis in multiple tissues. *PLoS Genet.* **9**, e1003486 (2013).
7. Grundberg, E. *et al.* Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.* **44**, 1084–1089 (2012).
8. Price, A. L. *et al.* Single-tissue and cross-tissue heritability of gene expression via identity-by-descent in related or unrelated individuals. *PLoS Genet.* **7**, e1001317 (2011).
9. Westra, H.-J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat. Genet.* **45**, 1238–1243 (2013).
10. Yao, C. *et al.* Integromic analysis of genetic variation and gene expression identifies networks for cardiovascular disease phenotypes. *Circulation*

11. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
12. Mitchell, T. J. & Beauchamp, J. J. Bayesian Variable Selection in Linear Regression. *Journal of the American Statistical Association* **83**, 1023–1032 (1988).
13. Groves, A. R., Beckmann, C. F., Smith, S. M. & Woolrich, M. W. Linked independent component analysis for multimodal data fusion. *Neuroimage* **54**, 2198–2217 (2011).
14. Groves, A. R. *et al.* Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *Neuroimage* **63**, 365–380 (2012).
15. Kolda, T. G. & Bader, B. W. Tensor Decompositions and Applications. *SIAM Review* **51**, 455–500 (2009).
16. Yener, B. *et al.* Multiway modeling and analysis in stem cell systems biology. *BMC Syst Biol* **2**, 1 (2008).
17. Hoff, P. D. Hierarchical multilinear models for multiway data. *Computational Statistics & Data Analysis* **55**, 530–543 (2011).
18. Khan, S. A., Leppaaho, E. & Kaski, S. Bayesian multi-tensor factorization. *arXiv.org* 1–23 (2014).
19. Buil, A. *et al.* Gene-gene and gene-environment interactions detected by transcriptome sequence analysis in twins. *Nat. Genet.* **47**, 88–91 (2015).
20. Brown, A. A. *et al.* Genetic interactions affecting human gene expression identified by variance association mapping. *Elife* **3**, e01381 (2014).
21. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. **348**, 648–660 (2015).
22. Reith, W., LeibundGut-Landmann, S. & Waldburger, J.-M. Regulation of MHC class II gene expression by the class II transactivator. *Nat. Rev. Immunol.* **5**, 793–806 (2005).
23. Stegle, O., Parts, L., Durbin, R. & Winn, J. A Bayesian framework to account for complex non-genetic factors in gene expression levels greatly increases power in eQTL studies. *PLoS Comput. Biol.* **6**, e1000770 (2010).
24. Kobayashi, K. S. & van den Elsen, P. J. NLRC5: a key regulator of MHC class I-dependent immune responses. *Nat. Rev. Immunol.* **12**, 813–820 (2012).
25. Pillai, R. S. *et al.* Unique Sm core structure of U7 snRNPs: assembly by a specialized SMN complex and the role of a new component, Lsm11, in histone RNA processing. *Genes Dev.* **17**, 2321–2333 (2003).
26. Liu, C. *et al.* MirSNP, a database of polymorphisms altering miRNA target sites, identifies miRNA-related SNPs in GWAS SNPs and eQTLs. *BMC Genomics* **13**, 661 (2012).
27. Melchjorsen, J. *et al.* Differential regulation of the OASL and OAS1 genes in response to viral infections. *J Interferon Cytokine Res* **29**, 199–207 (2009).
28. Potu, H., Sgorbissa, A. & Brancolini, C. Identification of USP18 as an important regulator of the susceptibility to IFN-alpha and drug-induced apoptosis. *Cancer Res.* **70**, 655–665 (2010).
29. Malakhova, O. A. *et al.* UBP43 is a novel regulator of interferon signaling independent of its ISG15 isopeptidase activity. *EMBO J.* **25**, 2358–2367 (2006).

- 546 30. François-Newton, V. *et al.* USP18-based negative feedback control is
547 induced by type I and type III interferons and specifically inactivates
548 interferon α response. *PLoS ONE* **6**, e22200 (2011).
- 549 31. Burkart, C. *et al.* Usp18 deficient mammary epithelial cells create an
550 antitumour environment driven by hypersensitivity to IFN- λ and elevated
551 secretion of Cxcl10. *EMBO Mol Med* **5**, 967–982 (2013).
- 552 32. Huan, T. *et al.* Genome-wide identification of microRNA expression
553 quantitative trait loci. *Nature Communications* **6**, 6601 (2015).
- 554 33. Lemire, M. *et al.* Long-range epigenetic regulation is conferred by genetic
555 variation located at thousands of independent loci. *Nature Communications*
556 **6**, 6326 (2015).
- 557 34. Small, K. S. *et al.* Identification of an imprinted master trans regulator at
558 the KLF14 locus related to multiple metabolic phenotypes. *Nat. Genet.* **43**,
559 561–564 (2011).
- 560 35. Fokoue, E. Stochastic determination of the intrinsic structure in Bayesian
561 factor analysis. *Technical Report, Statistical and Applied Mathematical*
562 *Sciences Institute* (2004).
- 563 36. Rotival, M. *et al.* Integrating genome-wide genetic variations and monocyte
564 expression data reveals trans-regulated gene modules in humans. *PLoS*
565 *Genet.* **7**, e1002367 (2011).
- 566 37. GTEx Consortium, Ardlie, K. G. & Dermitzakis, E. T. Human genomics. The
567 Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene
568 regulation in humans. *Science* **348**, 648–660 (2015).
- 569 38. Hawrylycz, M. J. *et al.* An anatomically comprehensive atlas of the adult
570 human brain transcriptome. *Nature* **489**, 391–399 (2012).
- 571 39. Witten, D. M., Tibshirani, R. & Hastie, T. A penalized matrix decomposition,
572 with applications to sparse principal components and canonical
573 correlation analysis. *Biostatistics* **10**, 515–534 (2009).
- 574 40. Sun, S. A survey of multi-view machine learning. *Neural Comput & Applic*
575 **23**, 2031–2038 (2013).
- 576 41. Chen, R. *et al.* Personal omics profiling reveals dynamic molecular and
577 medical phenotypes. *Cell* **148**, 1293–1307 (2012).
- 578
579

580

581 **Figure Legends**

582

583 **Figure 1 : Graphical representation of the method.** The gene expression data
584 tensor (top left) is decomposed into the product of an individual scores matrix, a
585 tissue scores matrix and a gene loadings matrix (top right). Columns of the
586 individual scores matrix are then used as phenotypes in a GWAS using SNP
587 genotypes (bottom left) in order to uncover genetic variation correlated with the
588 latent components.

589

590 **Figure 2 : MHC Class II regulation.** Figures **a** and **b** shows two components
591 identifying a similar network in different tissues. (Top left) GWAS with the
592 component's individual scores vector as a phenotype. (Top right) Boxplot of
593 individual scores stratified by genotypes at the lead GWAS SNP. Boxplots show
594 the median, upper and lower quartiles, with whiskers extending to either 1.5
595 times the interquartile range (IQR), or to the most extreme data point if this is
596 within 1.5 times IQR. (Bottom left) Gene loadings for the component. Only gene
597 loadings with a PIP>0.5 are shown. (Bottom right) Tissue scores vector for the
598 component shown as a barplot.

599 **Figure 3 : MHC Class I regulation.** (Top left) GWAS with the component's
600 individual scores vector as a phenotype. (Top right) Boxplot of individual scores
601 stratified by genotypes at the lead GWAS SNP rs289749. (Bottom left) Gene
602 loadings for the component. Only gene loadings with a PIP>0.5 are shown.
603 (Bottom right) Tissue scores vector for the component shown as a barplot.

604 **Figure 4 : Histone RNA processing.** (Top left) GWAS with the component's
605 individual scores vector as a phenotype. (Top right) Boxplot of individual scores
606 stratified by genotypes at the lead GWAS SNP rs6882616. (Bottom left) Gene
607 loadings for the component. Only gene loadings with a PIP>0.5 are shown.
608 (Bottom right) Tissue scores vector for the component shown as a barplot.

609 **Figure 5 : Type I Interferon Response.** (Top left) GWAS with the component's
610 individual scores vector as a phenotype. (Top right) Boxplot of individual scores
611 stratified by genotypes at the lead GWAS SNP rs2401506. (Bottom left) Gene
612 loadings for the component. Only gene loadings with a PIP>0.5 are shown.
613 (Bottom right) Tissue scores vector for the component shown as a barplot.

614 **Figure 6 : Zinc finger gene network.** (Top left) GWAS with the component's
615 individual scores vector as a phenotype. (Top right) Boxplots of individual scores
616 stratified by genotypes at the lead GWAS SNPs, rs17611866 and rs12630796.
617 (Bottom left) Gene loadings for the component, with zinc finger genes on chr 19
618 highlighted in purple. Only gene loadings with a PIP>0.5 are shown. (Bottom
619 right) Tissue scores vector for the component shown as a barplot.

Figure 7 : Multi-omics data integration. Graphical representation of a linked decomposition for several genomic assays. A matrix decomposition is applied to each data type. The matrix decompositions identify a different loadings matrix for each data type and a shared individual scores matrix.

Online methods

Bayesian Sparse Tensor Decomposition Model

We use Y to denote the 3D array or tensor containing pre-processed gene expression measurements. Y has dimensions $N \times L \times T$ where N is the number of individuals, L is the number of genes and T is the number of tissues. We model Y as follows

$$Y_{nlt} = \sum_{c=1}^C A_{nc} B_{tc} X_{cl} + \varepsilon_{nlt}$$

where C is the number of components (also called factors). A is an $N \times C$ matrix with the c^{th} column containing the individuals scores of the c^{th} component. B is a $T \times C$ matrix with the c^{th} column containing the tissue scores of the c^{th} component. X is a $C \times L$ matrix with the c^{th} row containing the gene loadings of the c^{th} component.

The error term is modeled as $\varepsilon_{nlt} \sim N(0, \lambda_{lt}^{-1})$ where λ_{lt} is the precision of the error term at the l^{th} gene in the t^{th} tissue.

We deal with missing samples for a given tissue by not including them in the model likelihood. We introduce an indicator variable I_{nt} that equals 1 when gene expression has been measured in tissue t for sample n and zero otherwise. The likelihood is then given by

$$P(Y|\Theta) = \prod_{n,l,t} P(Y_{nlt}|\Theta)^{I_{nt}}$$

where Θ is the vector of model parameters.

650 We fit this model in a Bayesian framework, and place priors on the entries of the
 651 matrices A, B, X and also the precisions λ_{lt} . A key prior is the one we place on the
 652 elements of the gene loadings matrix X . We wish to encourage sparsity in the
 653 rows of this matrix, so we use a hierarchical ‘spike and slab’ prior⁴² of the form

$$\begin{aligned} X_{cl} &\sim p_{cl} N(0, \beta_c^{-1}) + (1 - p_{cl}) \delta_0 \\ \beta_c &\sim \text{Gamma}(e, f) \\ p_{cl} &\sim \rho_c \text{Beta}(q, r) + (1 - \rho_c) \delta_0 \\ \rho_c &\sim \text{Beta}(s, z) \end{aligned}$$

655 For the purposes of making inference easier (see **Supplementary Note**) we use
 656 the equivalent factorization of the spike and slab distribution as $X_{cl} = W_{cl} S_{cl}$
 657 where

$$\begin{aligned} W_{cl} &\sim N(0, \beta_c^{-1}) \\ S_{cl} &\sim \text{Bernoulli}(p_{cl}) \end{aligned}$$

659 For the elements of A and B we use standard normal priors $A_{nc} \sim N(0, 1)$ and
 660 $B_{tc} \sim N(0, 1)$.

661 Model fitting

663 We fit this model using Variational Bayes (VB)⁴³, which approximates the
 664 posterior distribution $P(\Theta|Y) \approx Q(\Theta)$. The approach iteratively refines the
 665 estimate $Q(\Theta)$, by minimizing the Kullbeck-Lieber (KL) divergence between
 666 $Q(\theta)$ and $P(Y|\theta)$, or equivalently maximize the negative free energy. Once
 667 converged, $Q(\theta)$ can be used to approximate properties of the posterior
 668 distribution. The full details of the parameter factorization we use, the resulting
 669 VB update equations and details of parameter initialization are given in the
 670 **Supplementary Note**. The resulting algorithm has complexity $O(NLTC^2)$ and
 671 can be run on a multi-core server. For the TwinsUK data analyzed in this paper
 672 the method took 20 hours for each of the 10 runs using 8 threads.

674

675 Our model has the ability to shrink an entire component to zero ($\rho_c = 0$) and
676 effectively remove that component from the model. In this way our model can
677 adaptively choose the number of components it needs. Just a small amount of
678 experimentation is needed to find a large enough value of C so that components
679 start being shrunk to 0. For the TwinsUK data we fit the model with 1,000
680 components and found that in all 10 runs of the method around 50 components
681 would always be estimated as 0.

682 Summarizing the Variational Bayesian posterior approximation

683

684 The form of the VB posterior for every entry of the gene loadings matrix X_{cl} has
685 the same spike and slab form as the prior. We use this distribution to calculate
686 the expected value, denoted $E_Q(X_{cl})$. We also calculate a Posterior Inclusion
687 Probability (PIP) that X_{cl} is not equal to zero, which is equal to $E_Q(S_{cl})$. We use
688 the PIPs to infer a network of genes for each component consisting of the genes
689 with a PIP > 0.5. We summarize the individual and tissue scores vectors in a
690 similar way by using the expected values of the VB posterior, $E_Q(A_{cl})$ and
691 $E_Q(B_{cl})$ respectively.

692

693 Identifying robust components

694

695 The model is complex and has a large number of parameters and there is no
696 guarantee that the VB algorithm will find a global solution when optimizing the
697 bound on the marginal likelihood. Running the method multiple times highlights
698 this issue. Some components are found consistently across multiple runs,
699 whereas other components only occur in a small number of runs. For example,
700 our method often uncovers components that show strong *cis* eQTL signals when
701 using the associated component scores as phenotypes. To identify robust
702 components, we implemented a method that clusters similar components across
703 different runs. We then focus on large clusters containing components from

704 multiple different runs, and use these as the basis for our search for novel
705 signals.

706

707 More specifically, we run our method 10 times and store the individual and
708 tissue scores, gene loadings and PIPs. We calculate the absolute correlation
709 between the individual scores for all pairs of components across the 10 runs.
710 Hierarchical clustering is then used to group components into clusters, using one
711 minus the absolute correlation as a dissimilarity measure. The clustering is
712 terminated when no correlations between clusters are above 0.6.

713

714 The components within each cluster are then combined. We take the mean of the
715 individual scores, tissue scores and gene loadings and the median PIPs. The
716 individual scores for each component cluster are then used as a phenotype
717 against a genome-wide dataset of SNPs on the same individuals to identify which
718 components have a genetic basis. We apply quantile normalization to the
719 individual scores before testing for association with SNPs. Tissue scores are
720 thresholded to obtain tissue activity patterns. The distribution of tissue scores
721 tends to be tri-modal with one, well defined mode centered on zero so a
722 threshold can easily be picked to set small score values to zero. We only test
723 averaged components calculated from clusters with a minimum (user-defined)
724 membership size, in order to focus on components that are robustly and
725 consistently identified across runs.

726

727 Analysis of the TwinsUK dataset

728

729 Gene expression levels were measured for 845 female twins from the TwinsUK
730 cohort using whole transcriptome sequencing (RNA-seq), with data in three
731 tissues (adipose, lymphoblastoid cell lines (LCLs) and skin) for the majority of
732 the individuals^{19,20}. Experiments were performed using the Illumina TruSeq
733 sample preparation kit and sequenced on a HiSeq2000 machine. Reads were
734 mapped on to the GRCh37 reference genome using BWA v0.5.9⁴⁴. Only reads that
735 map uniquely were used. We run the method using RPKMs (reads per kilobase
736 per million) after performing the following pre-processing and normalization

737 steps; (i) genes with >20% zeros in all three tissues are removed resulting in
738 18,409 genes, (ii) quantile normalization of expression data within each tissue,
739 (iii) rank based transformation of each gene onto a standard normal.

740

741 Samples were genotyped on a combination of the HumanHap300,
742 HumanHap610Q, 1M-Duo and 1.2MDuo Illumina arrays. Samples were imputed
743 using the 1000 Genomes Project Phase 1 reference panel (data freeze 10
744 November 2010) using IMPUTE2⁴⁵ and filtered (minor allele frequency (MAF) <
745 0.01 and IMPUTE info value < 0.8). Imputed genotypes were available on 795 of
746 the 845 individuals.

747 We also used 11 concurrently measured phenotypes that were available on the
748 samples (age, BMI, weight, height, total cholesterol, HDL cholesterol, LDL
749 cholesterol (calc), total triglycerides, adiponectin, insulin and glucose) and
750 variables derived from the sequencing. Specifically, we used (a) the mode of the
751 insert size calculated for each sample, which can vary between sequencing
752 library preps, (b) GC-content of the reads from a sample, which can vary due to
753 biochemical differences in library prep and lane effect, (c) date of sequencing
754 and (d) primer index.

755

756 We ran our method 10 times on the dataset and combined components across
757 runs via clustering (see above). **Supplementary Figure 39** shows the resulting
758 distribution of cluster size. Only those clusters with more than or equal to 5
759 components were then retained for GWAS.

760

761 We used a linear mixed model⁴⁶ to test an individual scores vector as a
762 phenotype against the SNP genotypes. The scores vector was subset down to the
763 795 individuals for which imputed genotype data was available. We used a
764 Bonferroni corrected significance threshold of 1×10^{-10} , calculated by scaling a
765 genome-wide significance threshold of 5×10^{-8} by 500 to account for the multiple
766 GWAS we perform.

767

768 Testing associations between individual scores vectors and phenotypes and
769 batch variables was also performed using a linear mixed model⁴⁶, again only
770 using 795 individuals. Only one member of each twin pair was used in the
771 associations with age. The categorical batch variables, date and primer index,
772 were dealt with by creating binary vectors (one for each category) and
773 individually using these as a fixed effect in the linear mixed model.

774

775 Gene Ontology analysis was carried out using the TopGO R package⁴⁷. Gene
776 ontology analysis evaluates whether a particular set of genes are enriched for a
777 GO term in comparison to a background gene set. TopGO uses Fisher's exact test
778 to get a p-value for enrichment based on the expected and observed number of
779 genes with a GO term. Of the 18,409 genes used in this analysis, 13,965 have GO
780 annotations. To get a significance level for this analysis we randomly sampled
781 10,000 sets of genes of a random size and performed an enrichment analysis on
782 each set. We take the smallest p-value from each gene set to create a null
783 distribution and use this distribution to estimate a significant level of 1%.

784

785 We use a linear mixed model⁴⁶ to perform direct associations between the SNPs
786 and the (normalized) expression levels of genes involved our components. In
787 order to account for unmeasured confounding factors, we fit the PEER model²³ to
788 each tissue's expression data with 15 factors and use these as covariates in the
789 mixed model. In addition to the PEER factors, we also include two phenotypes,
790 (age and BMI) and two tissue-specific batch variables (GC mean and insert size
791 mode) as covariates.

792

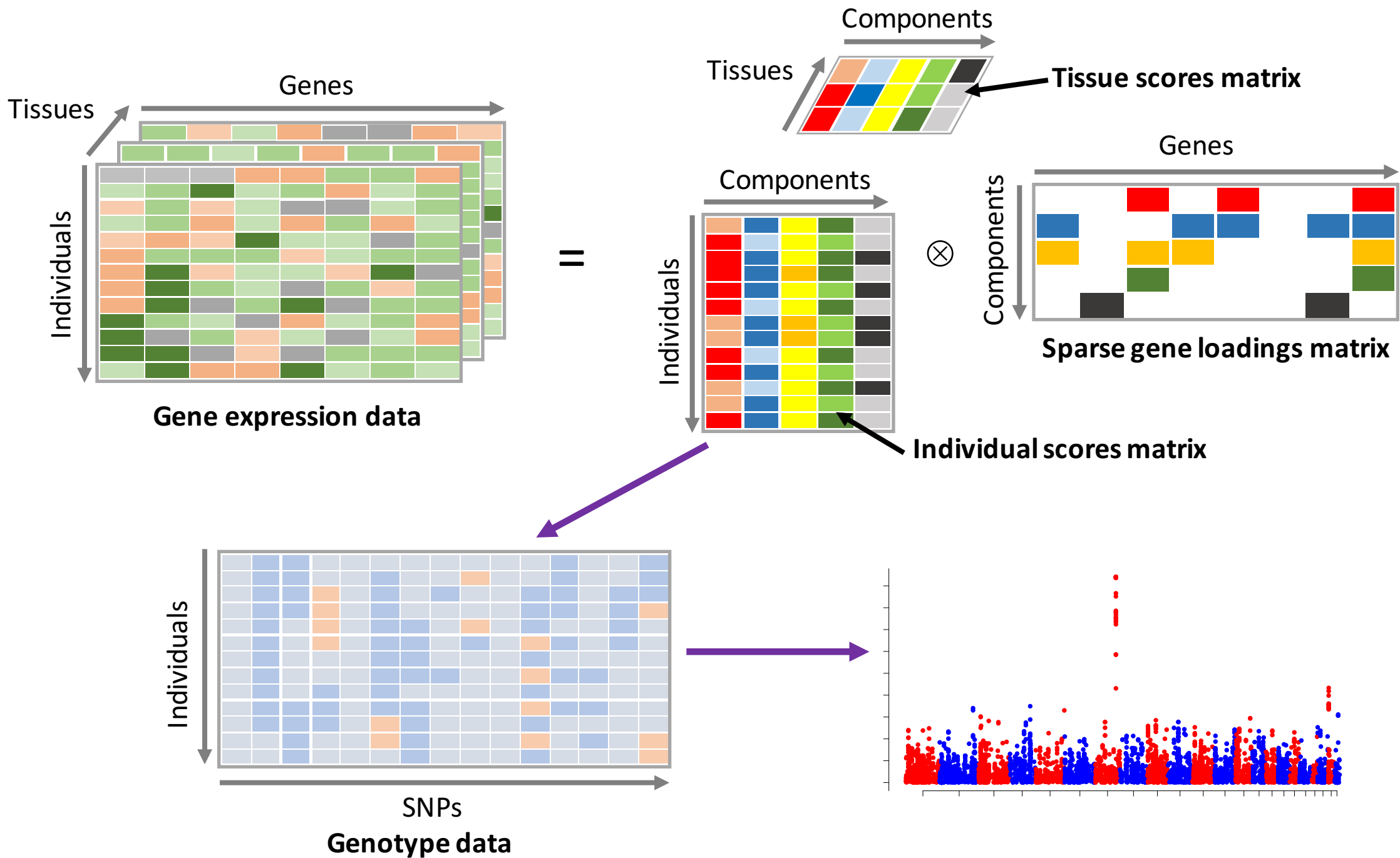
793 Application of fastICA

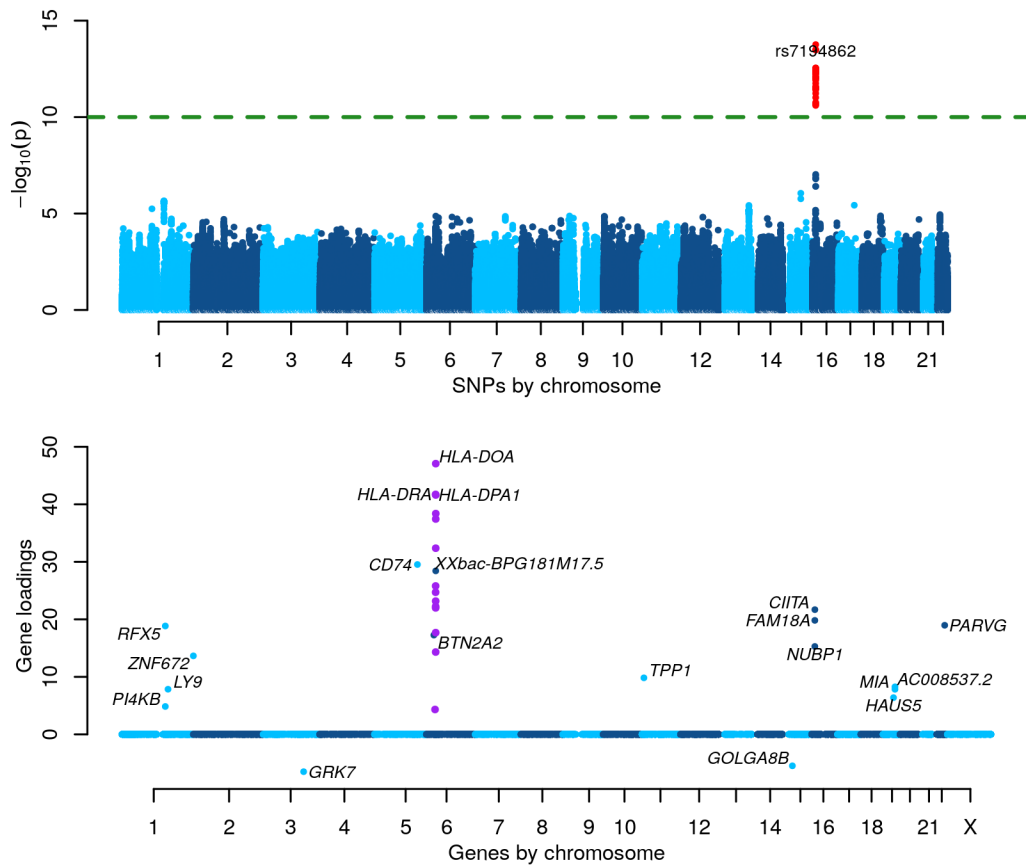
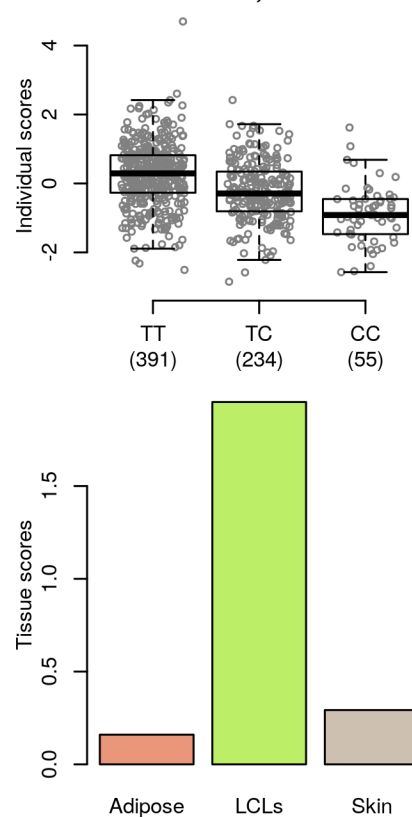
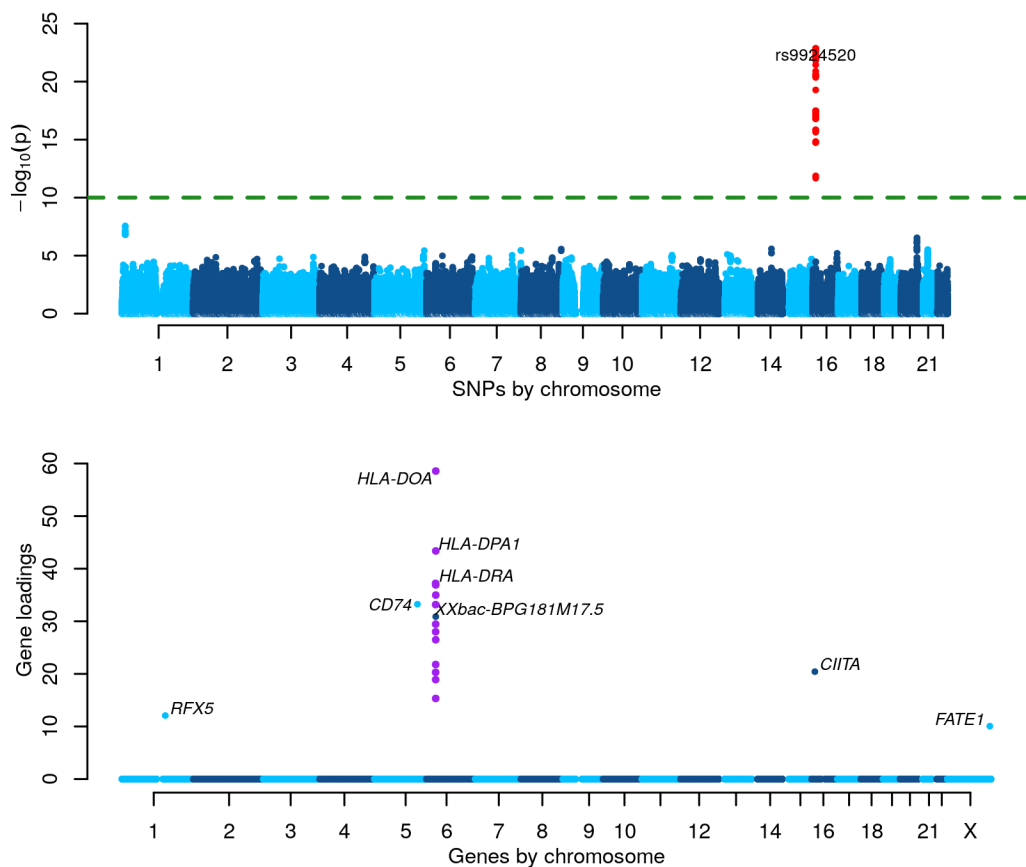
794 We used the R package fastICA to apply ICA to the TwinsUK dataset. We
795 concatenated the normalized expression data from the 3 tissues into a single
796 matrix. Only 618 out of 845 individuals had expression data on all 3 tissues, so
797 this matrix had 618 rows and 3×18409 columns. We fit the maximum number of
798 components possible (618). We selected the 200 components for the measure of
799 kurtosis of the gene loadings was > 3.5 and ran a GWAS against all SNPs. We also
800 tested the components individual scores against the known confounding

801 variables from the sequencing. More details are given in the **Supplementary**
802 **Note**.

803

- 804 42. Lucas, J. *et al.* in *Bayesian Inference for Gene Expression and Proteomics*
805 (eds. Do, K.-A., Muller, P. & Vannucci, M.) 1–25 (2006).
806 43. Jordan, M. I., Ghahramani, Z. & al, E. An introduction to variational methods
807 for graphical models. in 183–233 (MIT Press, 1999).
808 44. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-
809 Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
810 45. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands
811 of genomes. *G3 (Bethesda)* **1**, 457–470 (2011).
812 46. Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for
813 association studies. *Nat. Genet.* **44**, 821–824 (2012).
814 47. Alexa, A. & Rahnenfuhrer, J. *topGO: enrichment analysis for gene ontology*.
815 (R package version, 2010).
816



a**rs7194862, maf 0.282****b****rs9924520, maf 0.247**